Physics-constrained deep learning parameterizations for AGCMs

Katherine H. Breen, Donifan Barahona, Anton Darmenov, Matthew Thompson
NASA Global Modeling and Assimilation Office

Climate Dynamics Seminar
Dept. of Atmospheric, Oceanic, and Earth Sciences
George Mason University

March 9, 2022

Quick Overview

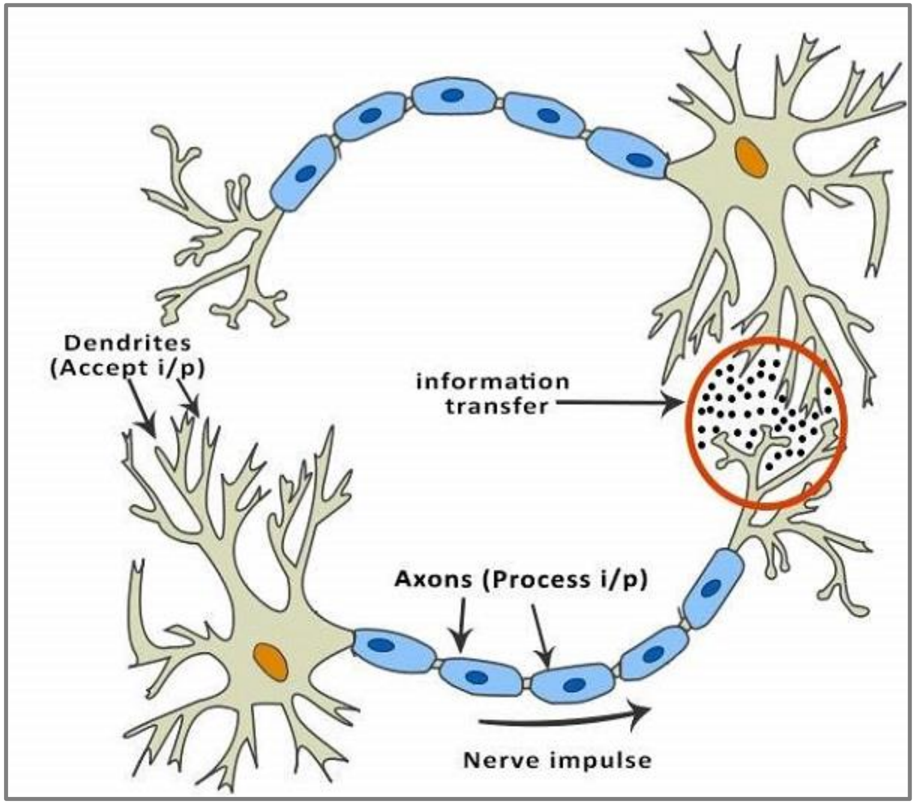
Donifan Barahona: Research Scientist, GMAO
Katherine Breen: NPP Fellow, GMAO

Research focus: Physics-based and AI hybrid modeling applications wrt cloud microphysics and aerosol-cloud interactions

Objectives:

- **Develop surrogate AI models for computationally expensive modules in GEOS that parameterize the contribution of clouds and aerosols to net radiative balance**
- **Use probabilistic modeling frameworks to assimilate sub-grid scale physics**

Artificial Neural Networks (ANN)



Anatomy of the brain:

Neuron A

Nerve/electrical impulse

Neuron B

Artificial neural network:

Input

How to change input to get desired output?

Output

General Workflow

Preprocess/ label data (normalization, one-hot encoding, etc.) → **Partition data into training, validation, and test sets** → **Train model on training/ validation sets**

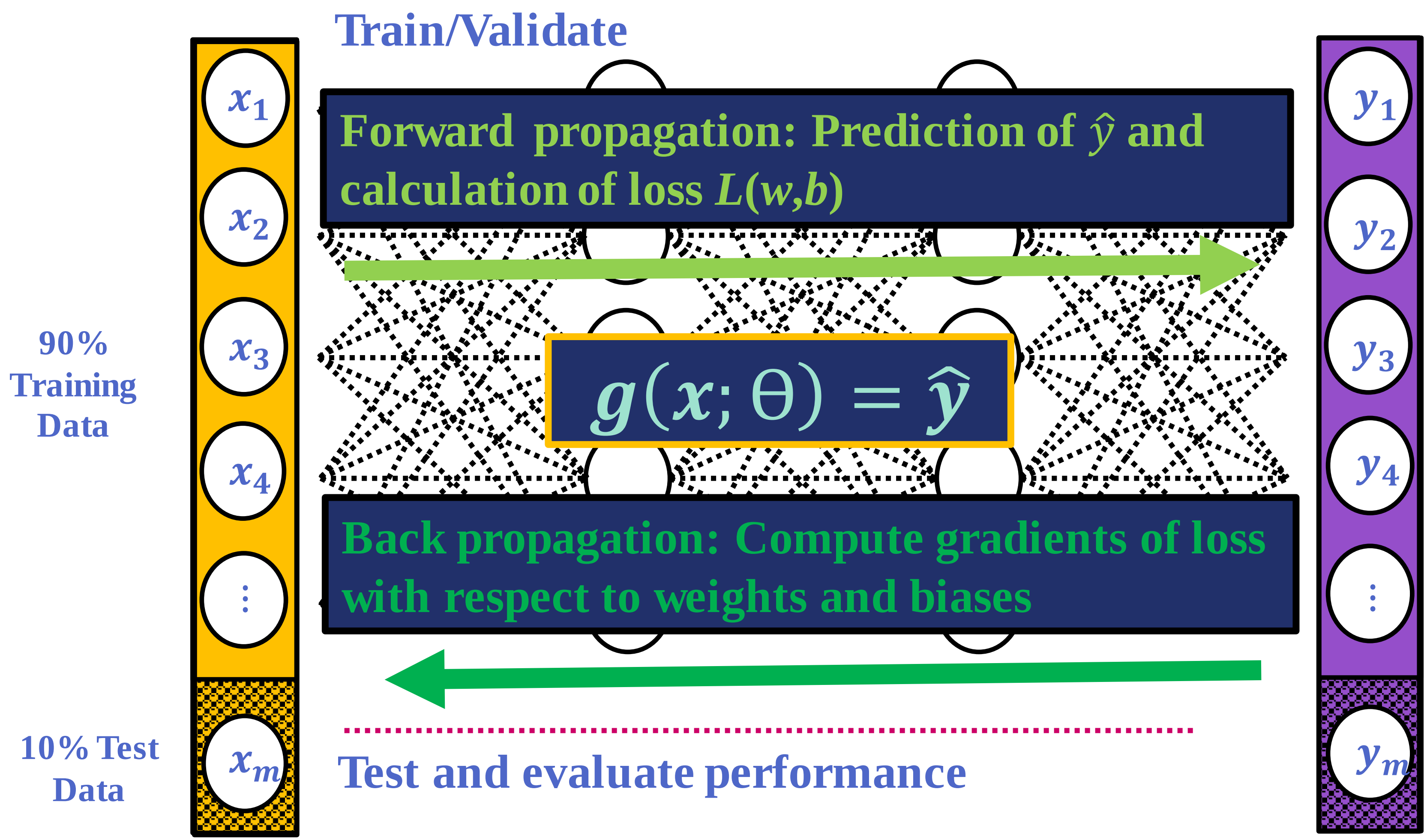
Predict on test

Training data – input/output used to train the model

Validation data – input/output used to evaluate the trained model on "unseen" data at regular intervals during training

Test data – input data used to predict output using the trained/validated model

How do neural networks learn?



What is an AGCM surrogate model?

Preprocessing – data-centric AI Data augmentation

Network architecture

AGCM inputs (X) → **Surrogate AGCM AI module** → **Prediction ($\hat{Y}$)** ↔ **AGCM outputs (Y)**

$\hat{Y} = \emptyset(\boldsymbol{W}\boldsymbol{X} + \boldsymbol{b})$
$\emptyset$ is an activation function,
$\boldsymbol{W}$ are learned weights, and
$\boldsymbol{b}$ are learned biases

Calculate loss ($\mathcal{L}(Y, \hat{Y})$), make updates for next training iteration

Probabilistic learning

Physics-guided loss function

Wnet: A Deep Learning-Based Parameterization to Predict Vertical Velocity *(W)*

1- Train a neural network (NN to predict p(W) (i.e., σ_W) using lower resolution state variables as input.

- Pros:
  - May capture important variability missed by a climatology
  - High resolution simulations provide a very large amount of data
  - Well-defined regression problem
  - Once trained it is relatively computationally inexpensive

2- Constrain the new NN parameterization by merging with observations

- Would bring unprecedented accuracy to the modeling of P(W)

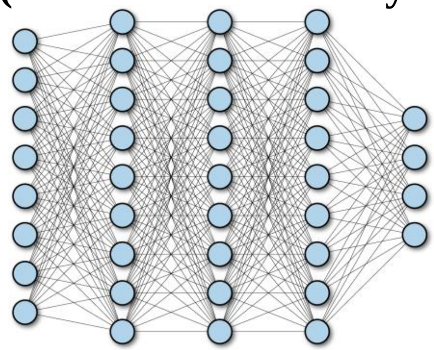
High Res GEOS 0.0625 °

Downsampling to 0.5 ° (mean, N=64)
U, V, T, Q…

Downsampling at 0.5° (stdev, N=64)
W

Training

“Wnet-prior”
(MLP ~ 5-6 hidden layers)



Transform

Transform and normalize
1 column = 1 sample

$\hat{Y}$

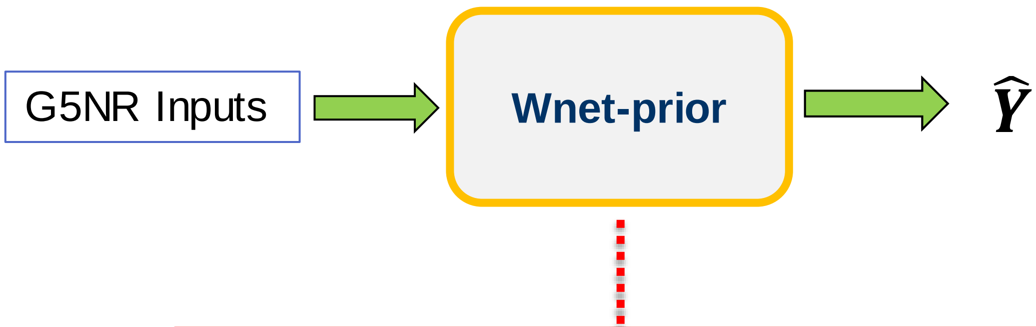
Calculate loss:
$\mathcal{L}(Y, \hat{Y})$

GEOS (online)

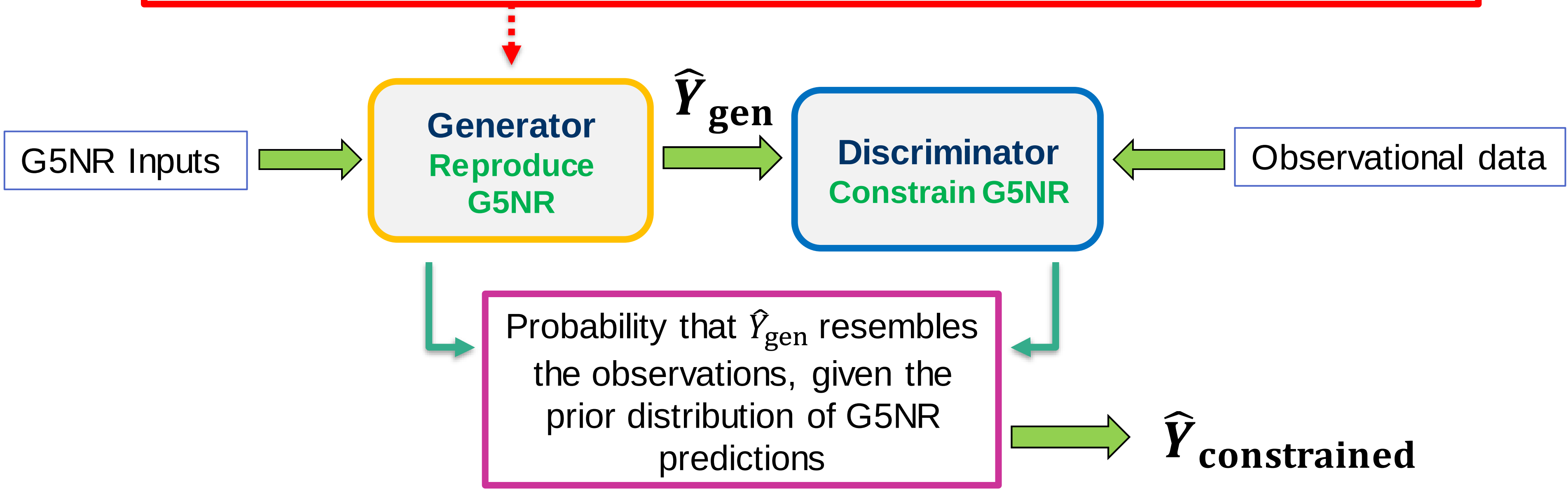
MERRA (offline)

Prediction $\hat{\sigma}_w$

Wnet-prior: Surrogate model for G5NR vertical wind velocity

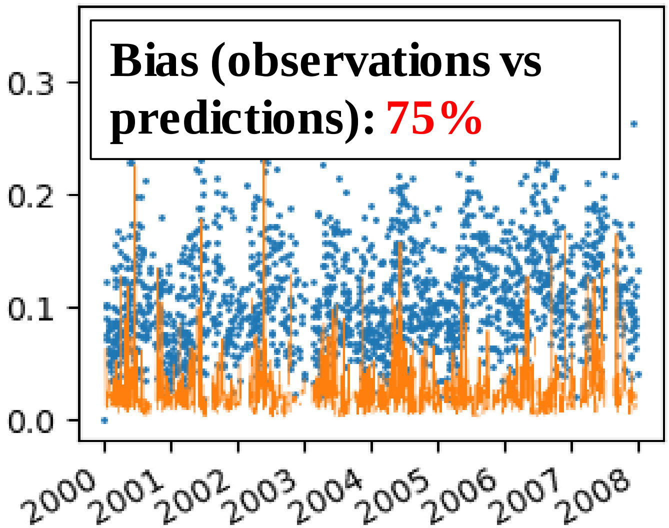


Wnet: Probabilistic refinement using observational data



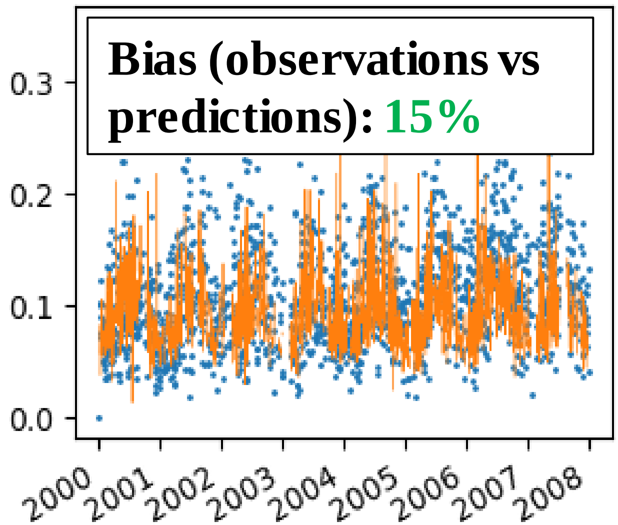
Wnet: G5NR surrogate for vertical wind velocity

W prediction (G5NR, unconstrained)



GAN

W prediction constrained by obs



Daily mean σ_w (m/s). Obs vs. – Prediction

AI brings unprecedented accuracy to parameterizations of sub-grid physics in GCMs

This approach can be extended to parameterize sub-grid variance in other variables/processes, e.g., total water and condensate, phase partitioning, aerosol emissions.

Summary

- Neural networks (NN) are useful to generate process-level cloud and aerosol parameterizations
- A NN parameterization (Wnet) was developed to estimate the subgrid distribution of vertical velocity using high resolution simulations. The NN was able to accurate reproduce the G5NR data.
- A novel GAN-based approach was developed to incorporate observations directly into the parameterization.
- Wnet is very versatile. It can be driven online during and GCM simulations or offline by reanalysis data MERRA-2.
- This approach could be extended parameterize subgrid variance in other variables/processes, e.g., total water and condensate, phase partitioning, aerosol emissions.

MAMnet: A Deep Learning-Based Parameterization to predict the Modal Aerosol Number Concentration

Train a neural network as a surrogate for the Modal Aerosol Module (MAM7) to predict modal number concentration

- **The Goddard Earth Observing System (GEOS) model consists of a set of components that numerically represent different aspects of the Earth system. The modal aerosol module (MAM7) was developed to simulate aerosol size distribution and number concentration, but at a high computational cost.**
- **The AGCM configuration of GEOS+MAM was used to develop a robust training dataset for a neural network (NN) surrogate model. We ran a 5-year GEOS+MAM simulation at 1-degree horizontal resolution with 72 vertical levels, generating ~1M samples per grid cell at 3-hourly temporal resolution.**

Why are aerosols important?

Accurate simulations of atmospheric chemistry require knowing the concentration of chemical species (aerosols) in the atmosphere

For a given species X:

- The **mixing ratio** (Cx) of X is the mass fraction of X (mass of X per mass of air)
- The **bulk number concentration** (Nx) of X is the number of molecules per unit volume of air.

Nx = Cx * Na => Nx = g(Cx) **We want to know g(Cx) for a size distribution!**

Where Na is the total number of molecules per unit volume air

How are aerosols represented in GEOS?
Two models: Goddard Chemistry Aerosol and Radiation model (GOCART) and Modal Aerosol Module (MAM)

	GOCART	MAM
Approach	Bulk – mass-based, single moment aerosol model, uses only total mass. All particles of the same size have the same composition.	Modal - two-moment aerosol model (predicts mass and number, infer size distribution per species)
Aerosol representation	Externally mixed (all species separate, all particles of the same size have same composition)	Internally mixed (multiple species in the same mode, more realistic)
Computational expense	Relatively low - **operational**	**High**
Notes	Comparable with assimilated aerosols (consistent with MERRA2)	Transport of additional tracers adds computational expense BUT **predicts particle size distribution – important for predicting radiation and aerosol-cloud interactions**

Objective

Develop a NN surrogate of MAM (MAMnet) to replace GOCART to predict modal aerosol number concentration (vs bulk)

TRAIN and VALIDATE: Identify a mapping between MAM mass mixing ratios and number concentration

MAM TRAIN/VAL INPUT → MAMnet → MAM TRAIN/VAL OUTPUT

TEST: Apply trained model on MAM inputs reserved from training set

MAM TEST INPUT → MAMnet → **Compare prediction with MAM test output data.** **Did the surrogate accurately learn the physical problem?**

APPLICATION: Apply trained model to GOCART inputs

GOCART INPUT → MAMnet → **Compare prediction with MAM.** **Is the model useful? Can it be applied to new data?**

MAM7 I/O Variables

INPUT VARIABLES

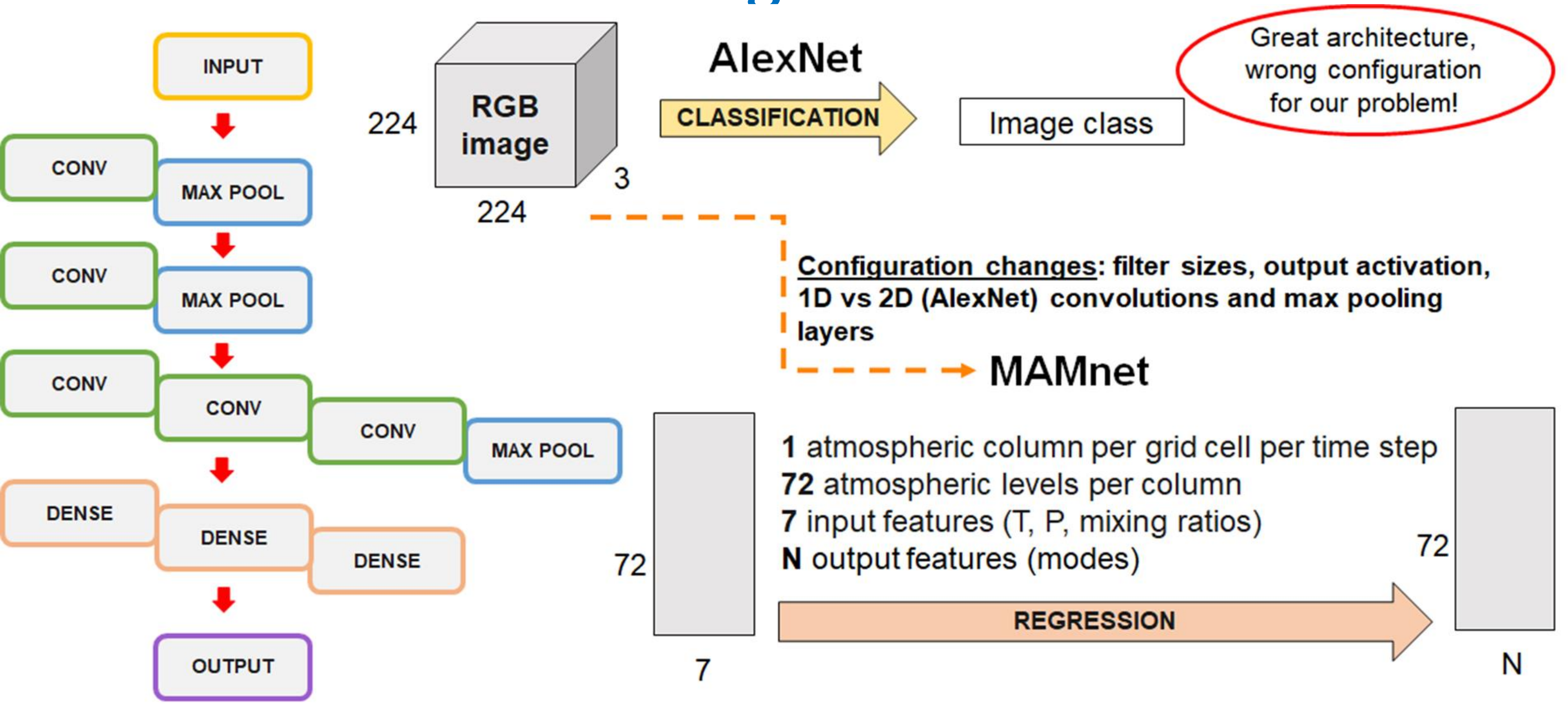
- Aerosol mixing ratios for:
  - Sulfate
  - Organics
  - Black carbon
  - Sea salt
  - Dust
- Temperature
- Pressure

→ **MAMnet** →

OUTPUT VARIABLES

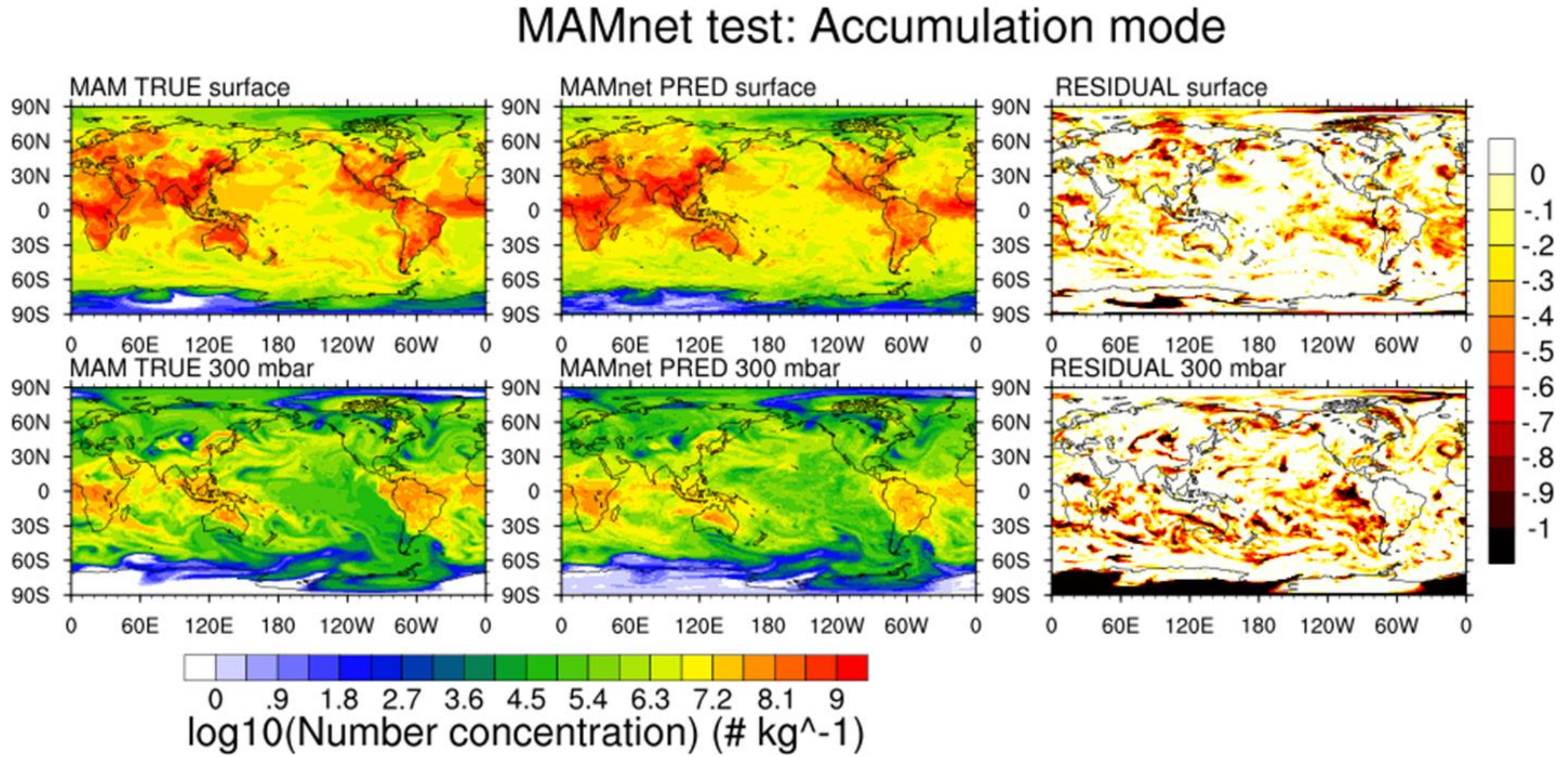
- Aerosol number concentration for 7 modes
  - Aitken
  - Accumulation
  - Coarse dust
  - Coarse sea salt
  - Fine dust
  - Fine sea salt
  - Organics

MAMnet Architecture: AlexNet adaptation for regression



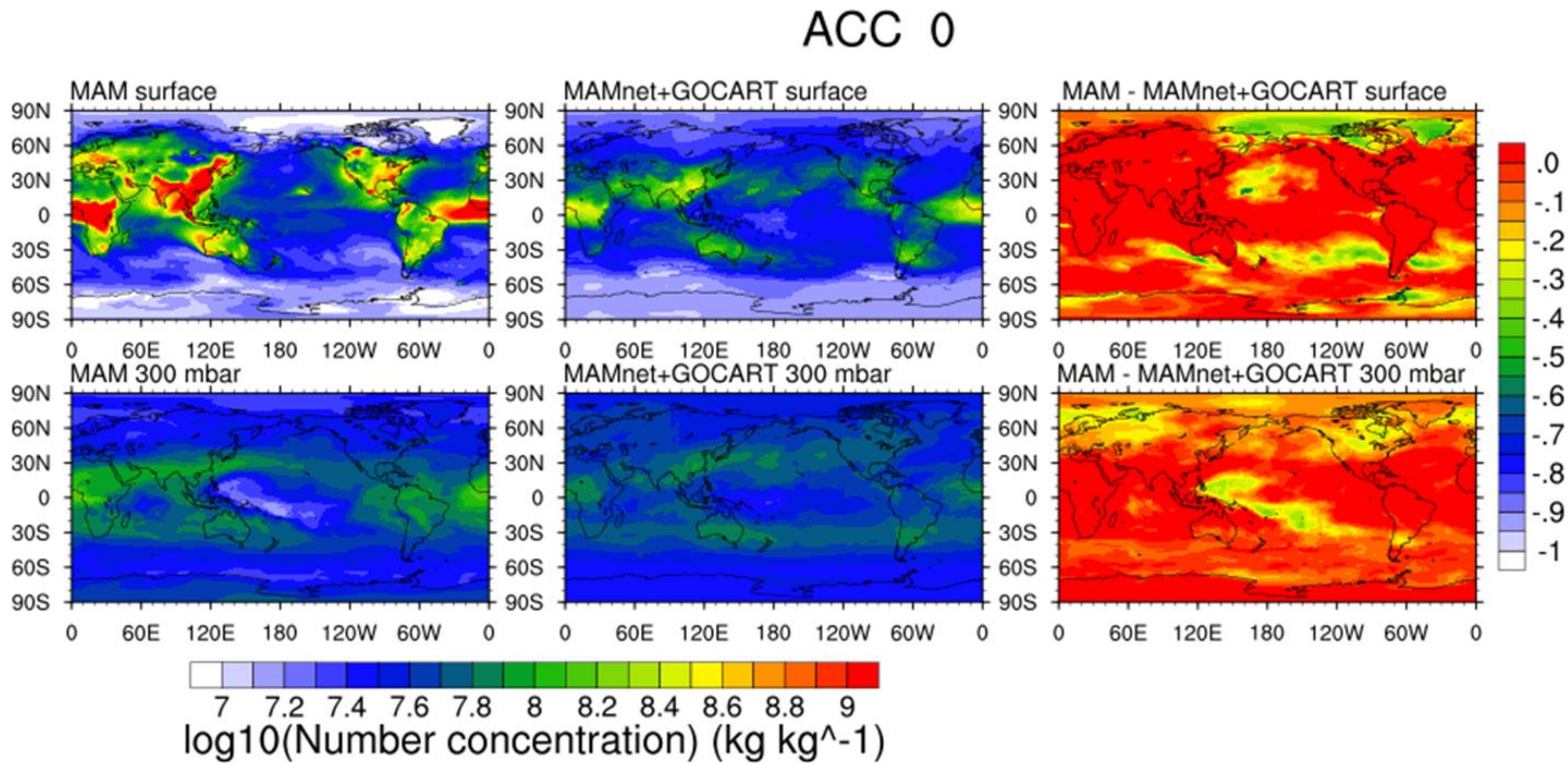
AlexNet citation:
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097-1105.

Did MAMnet accurately learn the physical Problem? YES



True and predicted data are aerosol number concentration in accumulation mode. Left: GEOS+MAM reserved test data ("true"). Middle: MAMnet prediction on GEOS+MAM test data. Right: Residual (true-pred). Top: surface. Bottom: Upper troposphere.

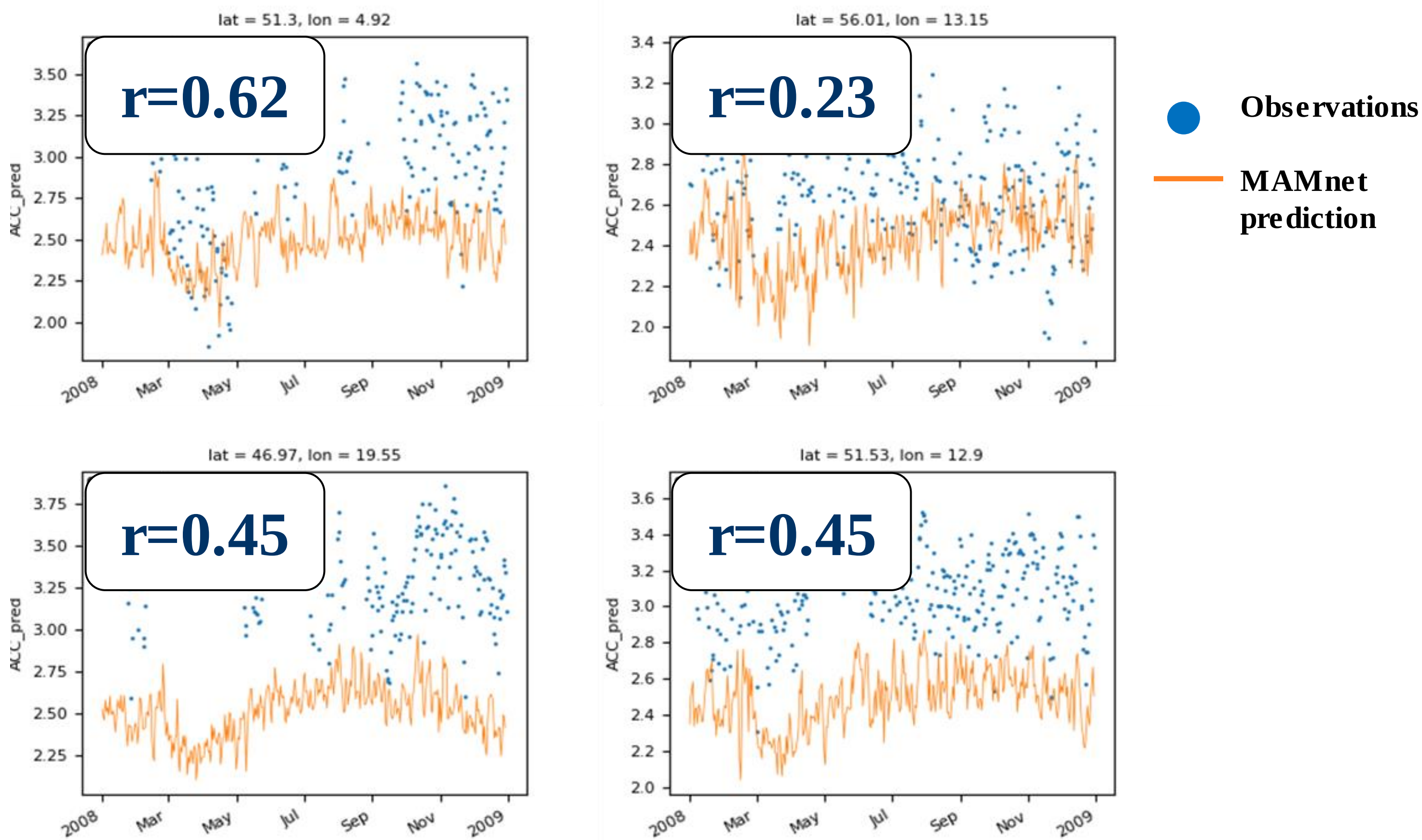
Is the model useful – can it be applied to new data? YES



Left: GEOS+MAM reserved test data ("true"). Middle: MAMnet prediction on GOCART inputs. Right: Residual (true-pred). GOCART is the operational aerosol number concentration module in GEOS.

Do MAMnet predictions compare well with observations?

YES



Trained MAMnet model run on reanalysis data (MERRA2; orange) compared with near-surface observational datasets (blue). See Asmi et. Al. (2011) for site descriptions.

Summary

Overall, MAMnet was able to reproduce the test data (not used during training/validation), indicating that MAMnet was trained on a dataset that encapsulated the statistical variability of the system and "learned" patterns from data as opposed to overfitting.

- **The AlexNet architecture was successfully adapted for a regression task on atmospheric data.**
- **These tests indicate that inputs from a low fidelity model (GOCART) were successfully used to replicate the output of a high fidelity model (MAM7) using a surrogate neural network**
- **Lastly, while MAMnet is able to reproduce observational data with some degree of accuracy, more development is necessary**